LARGE VOCABULARY SPEECH RECOGNITION IN FRENCH

Martine Adda-Decker, Gilles Adda, Jean-Luc Gauvain, Lori Lamel
Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{madda,gadda,gauvain,lamel}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this contribution we present some design considerations concerning our large vocabulary continuous speech recognition system in French.[1] The impact of the epoch of the text training material on lexical coverage, language model perplexity and recognition performance on newspaper texts is demonstrated. The effectiveness of larger vocabulary sizes and larger text training corpora for language modeling is investigated. French is a highly inflected language producing large lexical variety and a high homophone rate. About 30% of recognition errors are shown to be due to substitutions between inflected forms of a given root form. When word error rates are analysed as a function of word frequency, a significant increase in the error rate can be measured for frequency ranks above 5000.

INTRODUCTION

French speech recognition systems must address the high lexical variety of the French language which results in large out-of-vocabulary (OOV) rates. A large proportion of the observed lexical variety corresponds to homophones, which can be seperated only by an appropriate language model (LM). A comparative study of French and English showed that, given a perfect phonemic transcription, about 20% of the words in English newspaper texts are ambiguous, whereas 75% of the words in French newspaper texts have an ambiguous phonemic transcription[6]. Concerning lexical coverage, the number of distinct words in French must typically be double that of English in order to obtain the same word coverage under comparable conditions[6]. This difference between French and English mainly stems from the number and gender agreement in French for nouns, adjectives and past participles, and the high number of different verb forms[6]. This lexical variety can be partly reduced by appropriate text normalization [1], but there is a need for larger text corpora for training French LMs [2].

In this paper we address the impact of the text training data epoch and size on lexical coverage, language model (LM) perplexity and recognition results. Recognition results are presented and compared on 20k and 65k systems using test sets with and without control of the OOV rate. Word error rates are analyzed as a function of word frequency and root form normalization.

FRENCH RECOGNIZER EVALUATION

Some of our recent activities in LVCSR for the French language have been carried out in the context of a speech recognition evaluation project launched by the Francophone AUPELF-UREF organization. Academic sites with French recognition systems participated in various evaluation categories on read speech from *LeMonde* newspaper. The categories differed mainly by the allowed lexicon size (20k/65k), and by the use or not of an OOV-controlled test set. Previous experiments in LVCSR in French have been reported in [8] using a 20k vocabulary (LRE-SQALE project) on test sets with a controlled OOV rate of about 2%. Without artificial limitation the OOV rate tends to be closer to 5 or 6% with a 20k recognition vocabulary. For the AUPELF'97 evaluation [5], development and evaluation test sets containing about 180 paragraphs (600 sentences from 20 speakers) were selected without explicit control of the OOV rate. From this data ($\mathcal{T}$), a subset $\mathcal{T}'$, containing about 300 sentences were selected by including paragraphs with the lowest OOV rates.

The LIMSI system obtained the lowest word error rate of 11.2% (official result produced by the organizer[5]) on the evaluation test set (600 sentences). The word error on the development test data was 12.7% using the same system.

SYSTEM OVERVIEW

The recognition system configuration is extensively described in [4]. The acoustic parameters consist of 39 cepstral parameters (including first and second order derivatives) derived from a Mel spectrum estimated on a 8kHz bandwidth. Each acoustic model is a 3-state left-to-right CDHMM representing a phone in context. Gender-dependent models were trained using 66.5k sentences from 120 speakers of the BREF corpus[9]. For language modeling, 65k bigram and trigram LMs were trained on 205M words of *Le Monde* and *Le Monde Diplomatique* texts (years 1987-1996), and 64M words from *Agence France Presse* (AFP, years 1994-1996, distributed by LDC). Canonical pronunciations for the lexical entries were automatically generated using grapheme-to-phoneme rules [11]. These pronunciations were verified and

[1] Part of this work has been carried out within the ARC *Linguistics, Computer Sciences, and Spoken Corpora* supported by the AUPELF-UREF. The AUPELF-UREF is partially sponsored by the French government. ARC: Actions de Recherche Concertées, Coordinated Research Actions, AUPELF: Association des Universités Partiellement ou Entièrement de Langue Française, UREF: Université des Réseaux d'Expression Française.

alternative pronunciations were added semi-automatically. Each lexical entry is represented using a set of 35 phonemes.

Decoding is carried out in 3 passes: The first pass uses a small bigram LM (2.2M bigrams) to generate a word graph. The acoustic models used in this pass consist of about 3000 position-dependent triphones with about 8000 tied states. The second decoding pass, makes use of the word graph and a trigram LM (14M bigrams and 22M trigrams), and position-independent triphone models (about 9000 tied states distributed over 5000 models). In the third decoding pass unsupervised acoustic model adaptation based on MLLR [10] is carried out using the hypotheses generated in the second pass.

DESIGN CONSIDERATIONS

Previous work on LVCSR in French recognition [6, 8, 13] has highlighted the importance of increasing lexical coverage as an important issue in recognizer development. The link between coverage and language modeling is investigated more deeply here.

Lexical Coverage

The problem of lexical coverage has been addressed along different axes: word list size, word definition and word list selection. Text training corpora from *Le Monde* have been divided in different subsets[1] in order to assess the impact of training data size and epoch on vocabulary design:

T_0 : years 1987-88 (40M words)[2]
T_0' : years 1994-95 (40M words)
T_1 : years 1987-95 (185M words)
T_2 : years 1991-95 (105M words)[3]

Word list size: Better lexical coverage is obtained by increasing the number of words in the recognition word list. OOV rates are displayed in Table 1 for lexicon sizes ranging from 20k to 65k works, containing the N most frequent words in T_0 training data. The OOV rate decreases from over 6% to less than 2% on the development data.

word list	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
% OOV	6.4	4.3	3.2	2.4	2.0	1.8

Table 1: OOV rates on the AUPELF development set $\mathcal{T}$, for word lists ranging from 20k to 65k words. The word lists contain the N most frequent words in T_0 training data.

Word definition: For a fixed word list size, lexical coverage can be increased by applying appropriate language-dependent text normalizations. An extensive discussion of such text normalizations for French can be found in [1]. In Figure 1 OOV rates are shown to be reduced by about 50% when going from raw but clean data (text form N_a) to more aggressively normalized forms (N_b, N_c). The N_b form is derived from N_a by processing ambiguous punctuations, sentence-initial capitalizations, digits and acronyms. Form

[2] This was baseline resource for all partners in the AUPELF French recognizer evaluation project.

[3] T_2 is significantly smaller than T_1, but contains only the more recent data.

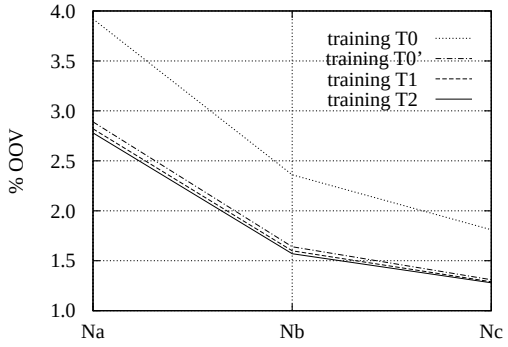

Figure 1: OOV rates on development test data for different normalization versions N_a, N_b, N_c on T_0, T_0', T_1, T_2 training data using 65k word lists.

N_c differs from N_b in that case sensitivity and diacritics are removed, and ambiguous punctuation markers are systematically decomposed. The N_b text form was used in this work.

Word list selection: A common approach for selecting the recognition vocabulary is to simply include the N most frequent words of the training texts. The selection of more representative training texts results in a better word list[7] as illustrated in Figure 1. Comparable lexical coverages are obtained for 40M, 105M and 185M word training text sets as long as they contain recent data (T_0', T_2, T_1). The use of 40M words of older data (T_0) entails a significant loss in coverage. This suggests that the training data epoch is more critical for optimal word list selection than the training data size.

Language Modeling

Text training material comprising a total of 255M words is taken from different sources:
LeM: 185M words from *Le Monde* years 87-96,[4]
MD: 6M words from *Le Monde Diplomatique*, years 89-96,
AFP: 64M words from *Agence France Presse*, years 94-96.

	LeM	*LeM + MD*	*LeM+ MD + AFP*
#words	185 M	191 M	255 M
#bg	11.9 M	12.1 M	13.5 M
#tg	13.9 M	14.3 M	18.1 M
ppx.	138	137	135

Table 2: LM size (#bigrams and #trigrams) and perplexity (ppx.) as a function of different training corpora: *LeM*, *LeM + MD*, *LeM + MD + AFP*. Bigram/trigram cutoffs of 0/1 respectively.

In Table 2 the LM sizes for fixed cutoff values are shown as a function of the training corpus size, along with the perplexity of the development data. When building trigram language models for French, we use smaller cutoffs (0/1) for bigram/trigram selection than we typically use for English (1/2). The lower cutoffs result in larger LMs, and suggest that still more data are necessary for accurate LM training.

[4] *LeM* corresponds to the T_1 corpus for lexical coverage.

Recognition Results

Recognition results with 20k and 65k systems on the AUPELF development set are shown in Table 3. The same acoustic model sets (described previously) are used for all conditions. The first two entries compare 20k systems with LMs estimated on the T_0 corpus (20k-40M) and and on the *LeM + MD + AFP* corpus (20k-255M). The third entry (65k-255M) corresponds to a 65k system where the output is filtered using the 20k vocabulary. The last entry (65k-255M) corresponds to the 65k system. For the 20k systems only small gains are observed despite the significant increase in the training text material with more recent data: 9% relative for the $\mathcal{T}$ condition and of 16% relative for the $\mathcal{T}'$ condition. Comparing the 20k-255M/20k and 65k-255M/20k systems the observed relative gain of about 25% for both $\mathcal{T}$ and $\mathcal{T}'$ can be attributed to LM improvements. With the same language model, an additional 20% relative error reduction is obtained by increasing the lexical coverage from 20k to 65k. These results illustrate the importance of increasing the system's vocabulary size provided appropriate LM training data are available.

		$\mathcal{T}$		$\mathcal{T}'$	
LM	Voc	%OOV	*%err*	%OOV	*%err*
20k-40M	20k	6.4	23.9	3.6	17.3
20k-255M	20k	6.4	21.8	3.6	14.6
65k-255M	20k	6.4	16.0	3.6	10.8
65k-255M	65k	1.3	12.9	0.5	8.8

Table 3: Recognition results with 20k and 65k systems on the AUPELF development set. The first column indicates the number of distinct lexical items in the LM, and the training text size. LMs estimated from the T_0 data or from the *LeM + MD + AFP* data. $\mathcal{T}$: 600 sentences, $\mathcal{T}'$: 300 sentence subset with controlled OOV rate.

To investigate the influence of LM training data size and epoch on recognition results, LMs were estimated from three different text corpora: T_0, T_0', and the 255M word corpus. LM perplexities and recognition results (without speaker adaptation) are reported in Table 4. A relative gains of over 10% is obtained by improving the epoch, with an additional gain of 8% relative by increasing the test size.

LM	ppx	*%err*
65k-40M (T_0, years 87-88)	198	16.8
65k-40M (T_0', years 94-95)	168	15.8
65k-255M (years 87-96)	135	14.5

Table 4: Results obtained with a 65k systems on the AUPELF development set $\mathcal{T}$. LMs are estimated from T_0 data, from T_0' data and from the *LeM + MD + AFP* data respectively.

Impact of OOV control

The last row of Table 3 shows the effect of controlling the OOV rate on the word error rate for a fixed vocabulary size. The error rate is reduced by over 4% absolute, which is more than 4 times the OOV reduction of 0.8%. This large difference is partially due to the difference of perplexities of the $\mathcal{T}$ texts (135) and the $\mathcal{T}'$ texts (106). This means that OOV control tends to filter out high perplexity sentences. Thus, the resulting test set is not only better covered by the word list, but also better modeled by the LM.

ERROR ANALYSIS

Recognition errors frequently involve incorrect gender, number and tense agreement and other homophone substitutions. A typical recognizer output is shown in Figure 5In or-

REF	Jacques Chirac pourrait il un jour remercier Alain Juppé et nommer Philippe Séguin au poste de premier ministre
HYP	Jacques Chirac pourrait il un jour remercié Alain Juppé est nommé Philippe Séguin au poste de premier ministre

Figure 2: Example of recognizer output illustrating most common error types: homophones: remercier $\rightarrow$ remercié (same root), nommer $\rightarrow$ nommé (same root), near homophones: et $\rightarrow$ est (word frequency rank < 50) .

der to further investigate the extent of these errors, the recognition error rates were compared for different test normalizations (see Table 6) using standard scoring. Starting from the N_b text form,[5] strings are normalized to remove case-sensitivity and diacritics, and decompounded (N_c form). Decompounding is the most effective normalization [3]. The importance of inflected form substitutions is shown by the two last entries in Table 6. Root forms were obtained using the INTEX system [12]. A relative error reduction of over 20% is obtained by replacing inflected forms by their root forms.

Normalization	*%err*
N_b form (baseline)	13.6%
N_c (N_b + no comp., ci, no diac.)	12.7%
N_b + root forms	10.3%
N_c + root forms	9.6%

Table 5: Word error rates as a function of different text normalizations applied to the N_b form of the reference transcripts and recognizer hypotheses. The two last entries of the table result from reducing inflected forms to root forms.

Word error rates are typically measured on individual sentences. In order to investigate how the word error rates are related to the LM accuracy, we propose to measure the word error as a function of the word frequency. To do so, the system vocabulary was partitioned into 11 word frequency rank regions (FRRs), $]K_{i-1}, K_i]$, logarithmically distributed along the decreasing word frequency axis. Each word w_n of the test set is associated its frequency rank k_n in the recognition vocabulary. If $k_n \in]K_{i-1}, K_i]$ then the FRR of w_n is K_i. The first FRR ($k_n \in]0, 10]$) contains the 10 most frequent words in the training data: *de, la, l', le, à, et, les, des, d', un* which are forms of defined and undefined articles, the conjunction *and*, and prepositions *of* and *to*. OOV words are grouped in an separate subset ($k_n > K_{11}$). Error rates can then be measured for each subset. Figure 2 shows the word error rate as a function of the 11 word frequency

[5] This form is slightly different from the official scoring normalization.

rank regions. The word occurrence distribution of the test data is provided in the same figure for reference. The OOV subset (1.3% of the data) with a 100% error rate is not represented. For each curve the figures are plotted at the upper bound of each FRR. For ranks $k_n > 5000$ error rates tend to increase drastically, but only concern about 15% of the test data, i.e. words not occurring in the first 7 rank regions. The first FRRs contain mostly short words (including monophone homophones) which are acoustically difficult to discriminate. The lowest word error is obtained for words in the 5th FRR (with frequency ranks from 500 to 1200). These words are well trained in the LM and they are usually polysyllabic and therefore acoustically easier to discriminate than words in the lower FRRs.

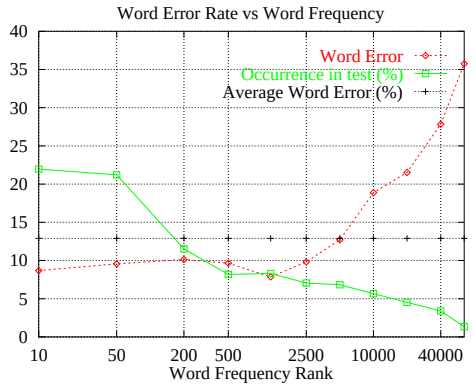

Figure 3: Word error rates and word occurrences as a function of frequency rank regions (FRRs) in the 65k system vocabulary. Each point defines the upper limit of an FRR. 11 FRRs are distributed logarithmically from 1 to 65000.

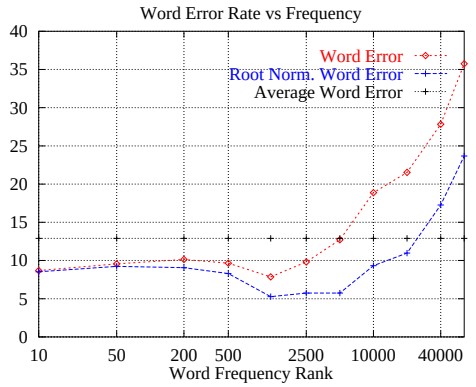

Figure 4: Word error rates (standard text and root form normalization) as a function of frequency rank regions (FRRs) in the 65k system vocabulary. The comparison of the two curves indicates the contribution of inflected form substitutions to the global error rate.

In Figure 3 word error rates on root form normalized reference and hypothesis strings are compared to the N_b form. For the less frequent words, substitutions between inflected forms of a given root form are seen to be an important source of error.

DISCUSSION AND PERSPECTIVES

Even though increasing the recognition vocabulary size is the most efficient way to reduce the OOV rate, small additional gains in lexical coverage can be obtained by optimizing the text normalization and weighting subsets of the training texts. For the latter the training text epoch is more important than the training text size. Increasing the training text material from 40M to 255M words allows larger LMs to be trained, which result in significant decreases in perplexity and in error rates.

The word recognition error was reduced by 40% (relative) by extending the vocabulary from 20k to 65k, and can be attributed to simultaneous improvements in coverage and language modeling. We have experimentally shown that OOV control has the side-effect of controlling the perplexity, which contributes to artificial performance improvements. Recognition errors are mainly due to homophones, for the most part errors in gender and number agreement. Error rates have been shown to increase dramatically for infrequent words, where an important rate of inflected form substitutions has been demonstrated. Improving language modeling techniques can be considered a most challenging research direction for French speech recognition.

REFERENCES

[1] G. Adda et al., "Text Normalization and Speech Recognition in French," *EuroSpeech'97*, Rhodos, Sept. 1997.

[2] M. Adda-Decker et al., "Elements pour la mise au point de système de reconnaissance grand vocabulaire en français," XXIIèmes JEP, Martigny, June 1998.

[3] M. Adda-Decker et al., "On the use of speech and text corpora for speech recognition in French," LREC'98.

[4] G. Adda et al., "Le système de dictée du LIMSI pour l'évaluation AUPELF'97," *1ères JST FRANCIL*, April 1997.

[5] J.M. Dolmazon et al., "ARC B1 - Organisation de la 1e campagne AUPELF pour l'évaluation des systèmes de dictée vocale," *1ères JST FRANCIL*, Avignon, April 1997.

[6] J.L. Gauvain et al., "Speaker-independent continuous speech dictation," Speech Communication **15**, pp. 21-37, Sept. 1994.

[7] J.L. Gauvain et al., "The LIMSI 1995 Hub3 System," *Proc. DARPA Speech Recognition Workshop*, Feb. 1996.

[8] L. Lamel et al., "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech'95*, Madrid, Sept. 1995.

[9] L.F. Lamel et al., "BREF, a Large Vocabulary Spoken Corpus for French," *EuroSpeech'91*, Genoa, Sept. 1991.

[10] C.J. Legetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**, pp. 171-185, 1995.

[11] P. Boula de Mareüil, "Pour une approche par règles en transcription graphème-phonème", Séminaire GDR-PRC CHM Lexique et Communication Parlée, Toulouse'96 France.

[12] M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, 1993.

[13] S.J. Young et al., "Multilingual large vocabulary speech recognition: the European SQALE project," *Computer Speech & Language*, **11**(1), pp. 73-89, Jan. 1997.